学校编码：10384　　分类号_____密级_____

学　　号：10520121152848　　UDC_____

厦门大学

硕 士 学 位 论 文

社交网络用户影响力测量中的同质性因素研究

Homophily and Influence of Users in the Social Network

岳豪

指导教师姓名：　苏俊斌副教授

专 业 名 称 ：　新　闻　学

论文提交日期：　2015 年 4 月

论文答辩日期：　2015 年 5 月

学位授予日期：　2015 年 6 月

答辩委员会主席：_____

评　　阅　　人：_____

2015 年 5 月 25 日

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

2015 年　月　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

2015 年　　月　　日

摘 要

社交网络中具有影响力的用户是积极的信息传播者，他们往往能够吸引大量的关注、转发与评论，设置议程并影响舆论导向。如何准确地找出具有影响力的用户并计算出社交网络中用户的影响力已经成为国内外学者共同关注的问题。目前大量的实证研究都是通过可见的、可测量的行为来衡量影响力的效果。然而在社交网络上具有同质性的个人、群体也会不约而同地采取一致的行为，这样就使得目前测量的用户影响力可能存在虚假成分。

社交网络上用户之间的同质性是一件非常复杂的事情。它可能是线上同质性，用户因为兴趣、性别、地理位置等相似而和陌生人建立联系；也可能是线下同质性，用户和好友因某些相似已在现实生活中建立了联系，社交网络只是他们联系方式的一种。如果社交网络上同质性显著的用户表现出一致的行为，那些将用户行为作为影响力评定维度的算法则需要重新调整公式，将同质效应剔除。

为了更全面地探究社交网络用户影响力测量中是否存在同质效应，本文将同质性分为线上同质性和线下同质性，以用户的转发行为、评论行为作为社交网络上用户行为的代表。线下同质性显著的用户是否具有相似的用户行为还需要继续探究。本文仅发现线下同质性显著的用户并没有表现出相似的的转发行为和评论行为，不能证实目前用户影响力的测量中因线下同质性存在虚假成分。

【关键词】社交网络； 影响力； 同质性

Abstract

Users with great influence in the social network are able to attract a lot of attention, influence other users forwarding or commenting on a piece of message. More and more scholars at home and abroad have paid much attention to how to measure the social influence of a particular node in a social network. While some algorithm didn't distinguish the real influence and homophily, the homogeneous users may have the same behaviours on social network which may exaggerate the influence of users. However, homogeneous users online may be strangers in real life who have online links due to some similarity, and it may also be offline homogeneous groups exerting their contacts online.The paper wants to explore whether two kinds of the homogeneous users will have the same behaviours such as forwarding and commenting on a piece of message.

【Key words】 social network；influence；homophily

目 录

Contents

第一章 导 言

1.1 研究背景

随着互联网技术的发展，越来越多不同形式、不同功能的社交网络开始走进人们的日常生活中并深刻地改变人们的生活方式。根据社交网站的功能，我们可以简单地将它们划分为交友网络（如 Facebook、Myspace 等）、媒体分享网络（如 Youtube、Flickr 等）、博客网络（如 Twitter、新浪微博等）、即时通信网络（如 MSN、QQ 等）和 BBS 论坛（Bulletin Board System，如天涯社区）等等（窦炳琳, et.al., 2012）。这些社交网络上庞大的用户数据为社会网络的研究提供了一个绝佳的平台.

中国互联网络信息中心的《2014 年中国社交类应用用户行为研究报告》表明，我国传统社交网站的覆盖率已达到 61.7%，新兴社交网站的覆盖率也达到 43.6%。

在社交网站如此高覆盖率的背景下，如何更好地衡量传播效果、如何准确地找出社交网络中具有影响力的用户已经成为国内外学者共同关注的问题。然而在社交网络上具有影响力的个体分布广泛，处于网络中的各个位置，且涉及到各个话题（丁兆云, et.al., 2011）。这在一定程度上给衡量社交网络用户影响力的算法带来了挑战。

目前衡量社交网络用户影响力的算法是在网络搜索等研究的基础上逐渐发展出的。以新浪微博为例，测量特定用户影响力的算法主要是基于特定用户被其他用户关注的数量(“粉丝数”)和用户与其他用户的互动行为（李军, et.al., 2012; Aiello, et.al., 2012)，例如用户发表的微博被其他用户回复、转发和评论等。 但此方向只从用户行为的某方面特征进行评估，并没有将用户本身的属性纳入考量，很难做到全面准确地评估用户的影响力。与此同时也有学者开始将用户的特性、话题的特性纳入考量，比如用户之间的同质性（Centola, 2010; Centola, 2011）、用户活跃度（杨长春, et.al., 2012）、话题的趣味性和情感倾向（Bakshy, et.al., 2011）

等。

衡量社交网络用户影响力最重要的也是最难的是将真正的影响力与混淆因素（Confounding Factors）区分出来。“这些混淆因素可能是同质性（Homophily），也可能是环境因素（Environment）” （Aral & Walker , 2012）。用户受到影响，在行为或状态（Profiles）上发生改变。这些改变后的结果在这些混淆因素的作用下同样可以发生”(Aiello, et.al., 2012）。

社会学家（例如 McPherson 等）早就注意到，存在同质性的一群行动者不约而同的行为与影响力产生的结果在行为上具有相同的表现。Aral 等（2009）在基于即时通信（Yahho!IM）互动与手机应用(Yahho!Go)被用户采纳的研究中发现，社会属性相近的个体更有可能不约而同地采纳新的手机应用，而这种情况在以前则被判定为影响的结果。Aral 等（2009）在这个案例中证明, 传统影响力的测量方法比实际影响力夸大了 3 倍到 7 倍。 Aral 等（2009）所揭示的同质化产生虚假影响的现象，究竟只是基于线下熟人延伸到线上互动的、具有私密性质的即时通信个案情形，还是在所有基于社交网络的互动中具有普遍意义的通则发现？人们通常采用的社交网络影响力评价办法是不是也存在类似的同质虚假影响效应？

相似的个体独立地做出相近的选择和行为，其结果往往会类似于个体在其他行动者（Actor）的影响下做出的选择和行为。然而“影响是一个通过节点之间的联接改变节点属性的过程，同质性是基于节点属性选择性建立联接的过程”（La Fond & Neville,2010）。从本质上讲，同质性并没有造成个体态度、行为等特质的改变，因而不是影响力。

无论是在现实世界还是在社交网络，个体行为的发生可能有多种原因。大部分的实证研究难以透过外部表象来判断其作用机制究竟是社会影响或是同质性。然而在社交网络研究中，这个迄今难辨真伪的“影响力”恰恰又经常被用做衡量传播效果的重要指标。

1.2 研究问题

影响力一词应用广泛，但在社交网络用户影响力的研究中，学者通常只是就

自己关注的用户特性、信息的特性等将其操作化，使其可测量（Ye & Wu, 2010; Bakshy, et.al., 2011）。目前还没有一个公认的定义来说明社交网络中的影响力究竟意味着什么。一般来说，当行动者以其他行动者为参照，改变自身的意见、态度或行为并使之与他行动者相似时，我们就可以认定社会影响发生了作用。

目前大量的实证研究都是通过可见的、可测量的行为来衡量影响力的效果(Weng, et.al., 2010；Yamaguchi, et.al., 2010；杨长春, et.al., 2012；李军, et.al., 2012；齐超, et.al., 2014)。然而具有同质性的人们往往也会采取一致的行为(McPherson, et.al., 2001）。那么，现在常见的社交网络用户影响力测量方法是否由于不能有效地识别同质效应，而导致其测量结果存在被夸大的成分？

1.3 研究意义

如今越来越多的媒体会在突发事件的报道中引用当事人的社交媒体作为信源（信息的发布者）。本研究通过探究社交网络用户影响力测量中是否存在同质效应在一定程度上有助于识别出社交网络上可靠的信源，帮助相关学者对现有的影响力评估算法提出改进。除此之外本文通过研究同质性用户的行为有助于更好地理解社交网络上信息的传播过程、用户的行为以及由于同质性聚集在一起的粉丝社群，有助于更有效地对网络舆情进行正确地引导和监督。

1.4 论文框架

本文的写作框架主要包括五个部分：

第一部分，导言。本章主要介绍研究的相关背景，确定具体的研究问题，介绍研究的目的和意义，并提出研究框架。

第二部分，文献综述。本章对国内和国际学界有关影响力和同质性的学术研究进行梳理，厘清相关研究的演进脉络，说明相关研究的创新和局限。

第三部分，研究方法。根据研究内容确定适合的方法，本文将分别采用实验法和问卷调查法来分别探究线上、线下同质性显著的用户是否具有一致的转发行为和评论行为。

第四部分，实验数据分析。研究一的实验并没有收集到具有说服力的数据。研究二总共收集了 202 份问卷，通过软件 SPSS19.0 对收集来的数据进行比例检验以及比例显著性检验。

第五部分，研究结论与局限。通过对实验数据进行分析得出实验结论，总结本文的局限，为用户影响力和同质性的未来研究提出相应的建议。

第二章 文献回顾

2.1 社交网络影节点的影响力测量

社会网络（Social Network）指的是“由有限的一组或几组行动者及限定他们的关系所组成的网络”（Hanneman & Riddle,2005）。

社交网络（Social Network Sites，简称 SNS）指的是基于互联网服务的一种社会网络。它们允许用户建立一个公开或半公开的个人档案，连接和其共享信息的其他用户，查看在这一平台上所有与自己建立了关系的用户（Boyd & Ellison,2010）。

Lazarsfeld, Berelson 和 Gaudet（1944）在《人民的选择》（The People’s Choice）以及后来 Lazarsfeld 和 Katz（1955）在《个人影响》（Personal Influence）等研究中，注意到媒体在产生影响效果方面的作用机制是间接的，常常需要通过意见领袖的人际影响来对受众产生作用，也就是“两级传播流”（Two-step Flow of Communication）假说。

如果说 Lazarsfeld 和 Katz 等学者的“两级传播流”假说开始着眼于人际交往中的影响力，那么 de Sola Pool 和 Kochen（1978）的研究则开始将政治科学中的影响力跟人际网络联系了起来，开创了从网络的角度衡量影响力的先河。他们认为“影响力在很大程度上取决于是否能够通过恰当的渠道与关键人物产生联系 。”

熟人之间的联系建构了社会网络，然而影响力的发生和传递并不局限于相熟的个体之间。社交网络结构的复杂性为影响力的解释提供了更多的可能性。[①] Lorrain 和 White（1971）曾提出 “结构等价”（Structural Equivalence）的概念，他们认为在社会网络中占有相似位置的角色在结构上都可以看作是等价的。Friedkin 和 Johnson（1997）认为，如果初始意见相异的个体在社交网络中占据

①利用社交网络结构的复杂性解释影响力的文献参考了陈昭晖（2015）的《关于社会影响力的文献综述》中的部分文献。具体参考的文献如下：de Sola Pool & Kochen,1978;Lorrain & White,1971;Friedkin & Johnson,1997.

的位置等价，那么他们的意见会在社会影响的作用下逐渐趋同。

2.1.1 社交网络影响力的定义

在社交网络用户影响力的研究中，学者通常只是就自己关注的用户特性、内容特性等将其操作化，使其可测量（Ye & Wu, 2010;Bakshy, et.al., 2011）。目前还没有一个公认的定义来说明在社交网络信息的扩散中影响力究竟意味着什么（Cha, et.al., 2009；Yamaguchi,, et.al., 2010；Weng, et.al., 2010）。

有的学者采用了社交影响（Social Influence）的概念(Aral, et.al., 2009）。社交影响是指用户与他人交往的过程中改变他人的思想、行为或者其它性质特征并使之相似于本身相应特征的一种能力(Marsden & Friedkin,1993）。

有的学者认为一个具有影响力的用户是指在自己的社交网络中具有权威性的用户（Weng, et.al. , 2010；Yamaguchi, et.al., 2010）。但是权威的定义是什么、具体表现在哪些维度上，他们并没有给予具体的说明。

针对微博这个社交网络，李军等（2012）认为“用户微博影响力可以表示为用户发布一则消息并使其传播到其它用户的能力”，如果将其量化则可表示为“某一用户与该用户发布的微博在所有用户和微博范围内的得分及排名。”

在衡量社交网络节点影响力的算法中，不同的学者关注的用户特性也不同。学者关注的用户特性可综合概括如下：粉丝数、回复行为、转发行为、提及行为、评论行为、用户的兴趣、活跃度、发表内容的趣味性、情感倾向及用户已有的影响力。

即使是相同的用户行为特性，在不同的研究中学者也会用不同的称呼。例如用户的粉丝数量，有学者称之为“入度影响力”（Cha, et.al., 2010），有学者将其称为“粉丝数量影响力”（Ye & Wu ,2010）。

2.1.2 社交网络影响力

在计算机科学领域，社交网络用户影响力的主要评价方法可以归纳为以下 4

种具有代表性的类型：基于 PageRank 的评价方法、基于用户行为权值的评价方法、基于 PageRank 和用户行为权值的评价方法、基于 URL 追踪的评价方法(李军, et.al., 2012；齐超, et.al., 2014）。

一、基于 PageRank 的评价方法

Page 等（1999）曾提出利用网页之间链接关系确定网页的重要性。将网页之间的链接关系视为拓扑结构②（Topology Structure）中的边（Edges），此网页指向其它网页的链接是此节点的出度（Outdegree），指向此网页的链接是此节点的入度（Indegree）。如果一个节点的入度越大，那么它的重要性越高。在所有的指向节点中，越重要的节点权值也会更高。数学公式描述如下：

$$R(u) = c\sum_{v\in B_u}\frac{R(v)}{N_v}$$

图 2-1 PageRank 的计算公式

其中 u 和 v 表示两个不同的网页，R(u) 和 R(v) 分别表示 u 和 v 的 PR 值，B(u) 为所有指向网页 u 的网页集合，N(v) 为网页 v 指向的网页数量，C 为保证 Web 页的 PR 值是常数的规范化因子（Normalization Factor）。此算法的缺陷在于页面的 PR 值是均匀地传递到链出的页面上去的，并没有考虑页面本身的重要程度。

结合社交网络的特性，社交网络用户影响力的衡量与网页重要性的衡量在某方面是相似的。“如果用户的粉丝是具有较高影响力的权威用户，该用户也会是权威用户。该用户对粉丝的影响力由粉丝从此用户接受的信息量的多少来决定。这样的相似性使学者在测量用户影响力时常以 PageRank 算法做基础。”（Weng, et.al. , 2010）。

在 PageRank 的基础上，Weng 等（2010）将用户的兴趣纳入考量，提出了 TwitterRank 的算法。其基本思想是在社交网络中，个体对不同话题的影响力是不同的，个体最终的影响力等于其在各个话题上的影响力总和。

② 是指网络中各个节点相互连接的形式。现在最主要的拓扑结构有总线型拓扑、星形拓扑、环形拓扑、树形拓扑以及它们的混合型。

杨长春等（2012）在 PageRank 的基础上与用户的多种特性结合，提出了一种新的中文微博社区博主影响力排名（Influence Rank）算法模型。他们将用户活跃度、博文质量以及博主传播能力纳入考量。用户活跃度定义为博主平均每日发博数量，微博质量系数定义为平均每篇博文被转发和评论的次数，博主传播能力则等于每篇博文的质量系数与博主平均每天发博数相乘后的数字。

二、 基于用户行为权值的评价方法

Cha 等（2010）关注了社交网络上用户的关注（Follow）、转发行为（Retweet）和提及行为（Mention），并认为这三种行为分别代表了用户影响力的不同方面。举例说明，用户看到一个感兴趣的话题，就会关注发布此话题的用户，并将此话题转发，让关注自己的用户也可以看到此话题。当一个人发现或者自己撰写一条有意思的消息之后，就可以选择提及（“@”）别人, 希望某些人也关注这条消息。作者认为“用户粉丝数量的多少可以体现一个人的受欢迎程度。用户话题的转发是由话题的价值决定的，转发数量的多少可体现用户传播信息和观点的能力。用户提及他人的行为是由被提及的用户名字的价值决定的，用户提及他人的次数体现了用户让别人也参与到对话中的能力”（Cha, et.al. , 2010）。

作者通过 Twitter 的应用程序编程接口 (Application Programming Interface，简称 api）收取了所有用户的信息。此信息包括用户发布的 Twitter 和用户与其它用户的连接状况（Social Links），在处理数据的过程中将发布 Twitter 的数量少于 10 条或用户名字无效的（Invalid）的数据排除在外。

作者分别按照前述三种行为进行影响力测量，再将所得数值通过斯皮尔曼等级相关系数（Spearman's Rank Correlation Coefficient）进行两两比较得出研究结论。除此之外作者还探究了用户的影响力是否因时间的推移和话题类型的变化而有所改变。

斯皮尔曼等级相关系数是衡量两个随机变量依赖性的非参数指标（Zar, 1972）。假设有两个或两组随机变量分别为 X、Y，它们的元素个数均为 N，这两个随机变量的第 i 个值分别用 Xi、Yi 表示，1<=i<=N。其数学公式如下图所示：

$$\rho = 1 - \frac{6\sum (x_i - y_i)^2}{n(n^2 - 1)}$$

图 2-2 斯皮尔曼等级相关系数